Research on Data Sharing Model Based on Cluster

Xiaobin Qiu, Hongqian Chen, Nan Zhou

Research on Data Sharing Model Based on Cluster

Xiaobin Qiu, Hongqian Chen, Nan Zhou
Network Center, China Agriculture University, Beijing 100083, P.R. China
Qxb@cau.edu.cn

Abstract: The data center is the core of digital campus, and the quality of data center depends on the accessing efficiency of data sharing. In this paper we study the advantage and disadvantage of the three current popular sharing model, and propose a sharing model based on cluster according to the reality of the digital campus. By the contrast and analysis of performance, the result show that the model can meet the requirements of the data sharing well.

Keywords: Digital Campus, P2P, Data Warehouse, Middleware, HBase

1 Introduction

With the rapid development of information technology, the departments of many universities have established their own professional management system, such as management systems for Office Automation(OA), financial information, educational information, scientific research, books and so on. These systems run independently and bring great convenience to people. However, there is insufficient information island, data storage and other problems. This page will discuss the data sharing model that is competent to be applied to the digital campus.

2 Data Share Model

At present, there are three main ways to realize the data sharing, i.e. data warehouse model, sharing middleware model and P2P data sharing model.

2.1 Data warehouse model

The data warehouse gets data through dada extraction, cleaning, conversion and loading. And it describes the data of heterogeneous database system through the global schema. The data that shares in different systems stores in the data warehouse. This provides unified data interface, service of data access and data analysis. The model of data integration and sharing based on data warehouse copy the heterogeneous data source to the designated data warehouse[1]. The system only access the copy of data of other system but not the original data. Sharing model as shown in Fig.1:

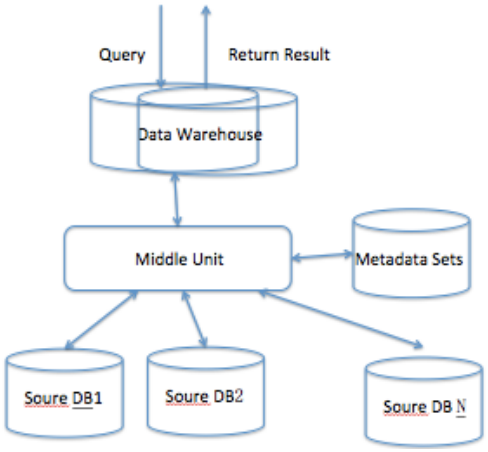

Fig.1. Data Sharing Model Based on Data Warehouse

The model connects the different systems to realize data sharing by data warehouse. The problems are:

(1) Quality. Data from other database by extraction, conversion, cleaning and other processing from the data source. Once the source database updated, the same semantic extraction data will be inconsistent and lead to data distortion.

(2) Reliability. Once there is the problem in the central database of centralized storage, the sharing data between different application systems will not work, and it will affect the whole integration and sharing system[2][3].

(3) Performance and Real-time. The data joins the data warehouse is by extraction, transformation, cleaning, loading operation and so on, it will take time too much and will affect efficiency of the system of high real-time requirements.

2.2 Middleware sharing model

The Middleware sharing mode is based on the global view mode. It realizes the data sharing among systems. The systems access each other by the middleware. There is a unified data logic view in the middle layer that manages the data sharing of source data. It can hide the detail of heterogeneous data to make the distributed source data into a logic whole. The middleware down coordinates the member databases and maintains the view mapping between the member database and middleware. And it up defines and specifies the transmission protocol and interface parameters. The model is focused to establish a global view to realize view mapping between the source data and the middleware that provide the perfect functions and reliable data service. The model shown in Fig.2:

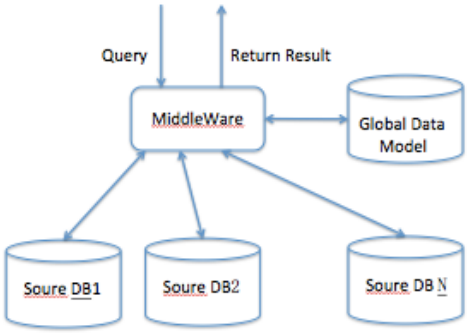

Fig.2. Data Sharing Model Based on MiddleWare

The middleware model can not only achieve transparent access to heterogeneous data source, but also the management is very simple and the security is good. The main problem is:

(1) The system failure rate increased. When system accesses the data of other system, as the middle layer, the number of links to access increases and the failure probability doubles that leads to the system failure rate increased.

(2)Efficiency. All sharing data transmits by middleware, and the system accesses frequently the middleware or transmits data by middleware that leads to high load. And the bottleneck is the middleware.

2.3 P2P Data Sharing Model

P2P(Peer to Peer Net) data sharing method is developed from the computing technology of P2P. The ideas are derived from the P2P network architecture. It is a distributed loosely coupled data sharing model that is developed from the traditional centralized data management. It accesses directly the database. This is equivalent to between two equal client or server data exchange. Each shares its data equally. The model shown in Fig.3:

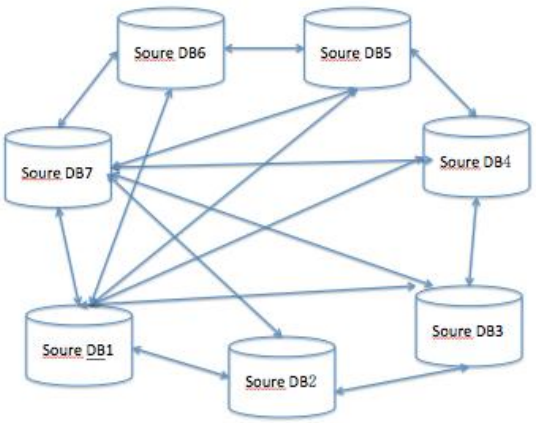

Fig.3. Data Sharing Model Based on P2P

P2P (peer-to-peer) data sharing has advantages of source data equivalence, flexible topology and distributed management, the main problems of the:

(1) Bottleneck of System. When the data in one node is shared by multiple system, and the number of accessed increases rapidly, the stress tolerance will be a great challenge, and the note will become the bottleneck.

(2) Node Failure. With the topology expanding and the number of nodes increasing, the complexity of system increases and the possibility of failure of a single node becomes larger. The failure of a core node will affect the operation of the whole system.

(3) Security flaw. Without QoS quality assurance, anonymous communication and data exchange is unsafe. Without Qos Quality assurance, delay, packet loss, availability and reliability will influence the quality of service.

According to the comparison of above, P2P and middleware tends to involve business system less or high real-time, and data warehouse model is more suitable for complex business. In this paper, according to the actual situation of the digital campus, and combining the advantages and disadvantages of the three kinds of model above, we proposed a data sharing model based on cluster[2][3].

3 3 Data Sharing Model Based on Cluster

The purpose of digital campus platform is to provide the service of data sharing to the department, and the more important purpose is to build a data warehouse facing theme, and then build a large data to statistic analysis, and provide the data basis for the leading. The connectivity operation and transaction characteristics of relation database, such as Atomicity, Consistency, Isolation, Durability, will make it poor efficiency in scalability and analysis capability. According to the principle of intensive construction of application system integration, the NoSQL database and cloud computing have advantages in data sharing in large-scale. It is high ability to handle large amounts of data and sharing, and it reference to the data warehouse storage sharing idea and concept of cloud storage for all kinds of data types such as structured, semi-structured and unstructured. The relation database is the most widely used before the appearance and development of NoSQL database in recent years. With the rapid processing requirements on large data, NoSQL database gradually shows its advantages. The NoSQL database is the distributed storage system of column oriented that can achieve high performance of concurrent read and write operations, such as Cassandra, Hbase, BigTable.

In this paper, we establish a data sharing model based on the cluster of public databases in a Hadoop HBase database as an example. In the model, it realizes the data sharing and the processing and analysis of data, but it can keep the original private business system. The model shown in Fig.4:

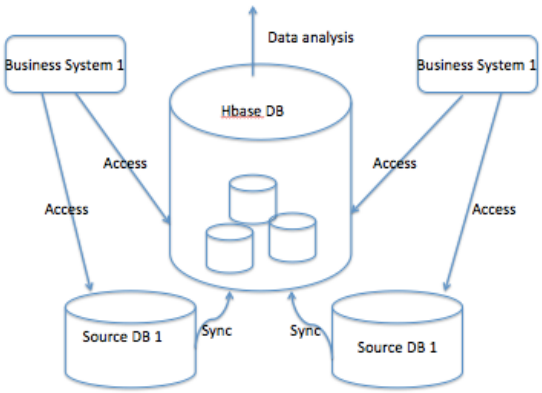

Fig.4. Data Sharing Model Based on HBase

In the model, the HBase database is the central data warehouse, and synchronizes periodically with the source database. The HBase stores multiple copies for public data to realize sharing efficiently. One system accesses the data of other systems by the mode of accessing public database directly. But the access mode should make corresponding adjustment because HBase is non relational database, and should access database by non relational database mode. This has advantage of the distributed parallel computing(Map/Reduce) and the distributed storage. The different businesses access the different sharing data through the access control[4][5].

4 Performance Analysis and Contrast

The data operation of sharing model can be divided into storage, transmission, query and so no. And the process includes extraction, cleaning, conversion, transmission, processing (add/delete/update), and the extraction, cleaning is a special steps of data synchronization. Now the data query as an example, the time required for each mode as:

Set:

Tc: Time of data converting

Tt: Time of data transmission

Tq: Time of data querying

The time of data warehouse model:

$T = Tt + Tq$ (1)

The time of middleware model:

$T = Tc + Tt + Tq$ (2)

Tc is the time of converting of middleware.

The time of P2P model:

$T = Tt + Tq$ (3)

The time of data sharing mode based on cluster:

$T = Tt + Tq$ (4)

The operation time is same but mode(2) that have Tc. The Tt is same in the same network, so the key is Tq. mode (1)(2)(3) use relation database and mode(4) uses non relation database.
The following is a group of test data. The key value is used as the query condition in order to ensure the consistency of data because of only key value query in HBase. In 5000000 of data, we query 1 to 100000. The test results in table 1:

Table 1. Query Speed Comparison of HBase and MySql

Number	HBase1	HBase2	HBase3	mysql1	mysql2	mysql3
1	619	602	638	2295	2190	2302
10	738	723	789	2909	2887	2931
100	1102	1124	1210	5232	5219	5214
1000	2522	2612	2460	10213	10207	10211
10000	5532	5510	5628	32012	31018	33101
100000	12092	12030	13021	52351	53121	53462

According to the test data, draw the diagram as Fig.5

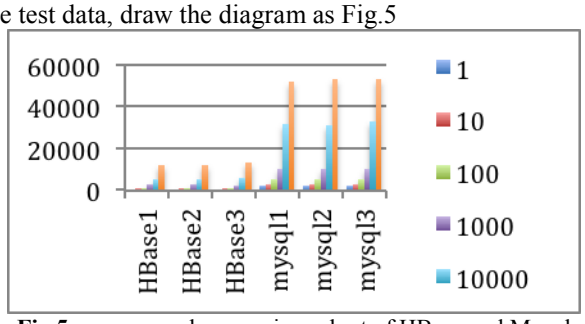

Fig.5. query speed comparison chart of HBase and Mysql

Our performance test results show that , the time of HBase and MySql does not increase obviously and, the time range is limited as the quantity of querying data increases,. However, when the volume of data reaches a certain amount, the performance advantage of HBase is obvious.
When the data quantity is growing, HBase can use the cheap PC server to realize the expansion. On the other hand, MySql may also be expanded, for example, multiple copies can be stored in different database servers, and each server has a copy of all the data, but each server can not load balancing automatic. because they serve individually for a part of the user.

5 Conclusion

This paper presents the advantage and disadvantage of the current data sharing model, then proposes a sharing model based on cluster. The performance test confirms that

this model has certain advantages. At present, the distributed data is used by many large company such as Google, Baidu, Alibaba and so on. With the development of the large data concept, the data sharing mechanism will be much further improved and widely used in the campus digital construction.

Reference

1.Mario Lassnig, Vincent Garonne, Gancho Dimitrov, LucaCanali.:ATLAS Data Management Accounting with Hadoop Pig and HBase. International Conference on Computing in High Energy and Nuclear Physics 1-7(2012)

2. Dave Dykstra.: Comparison of the Frontier Distributed Database Caching System to NoSQL Databases . International Conference on Computing in High Energy and Nuclear Physics 1-5(2012)

3.Craig Franke, Samuel Morin, Artem Chebotko, John Abraham, and Pearl Brazier.:Efficient Processing of Semantic Web Queries in HBase and MySQL Cluster. The IEEE Computer Society 36-41(2013)

4.Franke, Craig; Morin, Samuel; Chebotko, Artem.: Efficient Processing of Semantic Web Queries in HBase and MySQL Cluster. IT PROFESSIONAL 36-43(2013)

5.M. Kumar and C. Duffy.:An Object Oriented Shared Data Model for GIS and Distributed Hydrologic. International Journal of Geographical Information Science, IJGIS-2008-0131